

Method for Delivering Data Over a Network

Field of the Invention

This invention relates to methods and systems for delivering data over a network, particularly those for delivering a large amount of data with repetitive content to a large number of clients, like Video-on-Demand (VOD) systems.

Background of the Invention

Current VOD systems face a number of challenges. One of them is how to provide the clients, which may be in the number of millions, with sufficient interactivity like fast-forward/backward and/or forward/backward-jump. At the same time, the provision of such functions should not impose severe network load, as the network resources namely the bandwidth may be limited. Furthermore, every client generally prefers to have the movie he selects to be started as soon as possible.

The following sections describe some of the currently used VOD systems and their possible disadvantages:

1. Near-VOD (NVOD) with regular stream-interval

A NVOD system consists of staggered multicast streams with regular stream interval T (Figure 1). The streams are multiplexed onto the same or different physical media for distribution to the users via some multiplexing mechanisms (such as time-division multiplexing, frequency division multiplexing, code-division multiplexing, wavelength division multiplexing etc...). The distribution mechanisms include point-to-point, point-to-multipoint and other methods. Each stream is divided into regular segments of interval T , and the segments are labelled 1, 2, 3, ..., N respectively. The content that is to be distributed to the users is carried on the N segments and the content is replicated on all these streams. The content is also repeated on each stream in time. By using such a staggered streaming arrangement with regular stream interval T , the users are guaranteed to receive the content at any time with a start-up latency less than T . However, there is no provision for user interactivity in such a system. If a user interrupts the content viewing say by pausing the display, the user cannot resume

the viewing at the same play point where the user pauses and is forced to skip some content to keep up with the multicast-stream that is continuously playing.

2. Quasi-VOD (QVOD) with irregular stream-interval

5 A QVOD system consists of staggered multicast streams with irregular stream intervals (Figure 2). The streams are multiplexed onto the same or different physical media for distribution to the users via some multiplexing mechanisms (such as time-division multiplexing, frequency division multiplexing, code-division multiplexing, wavelength division multiplexing etc...). The distribution mechanisms include point-
10 to-point, point-to-multipoint and other methods. Unlike the NVOD system where the streams constantly exist, the streams in a QVOD system are created on demand from the users' request for the content. The users' requests within a certain time interval T_i are batched together and served together by Stream i . The stream intervals $T_1, T_2, \dots, T_i, \dots$ are irregular. The streams (Stream 1 to i etc...) are all provided on-demand and
15 will be removed as soon as the content distribution has been completed. The streams are constantly created as users' requests come in. By using such a staggered streaming arrangement with irregular stream interval T_i , the particular group of users starting within interval T_i is guaranteed to receive the contents within T_i (start-up latency). Again, there is no provision for user interactivity in such a system. If a user interrupts
20 the content viewing say by pausing the display, the user cannot resume the viewing at the same play point where the user pauses and is forced to skip some content to keep up with the multicast-stream that is continuously playing.

3. Distributed Interactive Network Architecture (DINA)

25 DINA system refers to the method and system as described in the applicant's PCT applications PCT/IB00/001857 & 001858. In the DINA system, interactive functions including fast-forward/backward, forward/backward-jump, slow motions, and so on can be provided by a plurality of multicast video data streams in conjunction with a plurality of distributed interactive servers. Although interactive
30 functions may be provided to the client in such the DINA system, the network load may increase if the start-up time for each user's request is to be reduced. This is determined by the stream interval of the multicast data streams. Generally, the number of data streams, and therefore the network load, increases with the decrease of the stream interval.

In the NVOD and QVOD systems, a user wanting to view the content will simply tap into one of the many staggered streams and view the content simultaneously with all others sharing the stream. While such schemes are simple and efficient, they suffer from two difficulties – a large start-up latency and user inflexibility.

For the first difficulty, a user may have to wait as long as one stream interval T before the request is served, and the waiting time may be as large as many minutes or even hours, depending on the stream interval. Although the stream interval can be made very small, say even down to a few seconds, this also means that the system has to provide a large number of streams for serving the same amount of content. The number of streams required is simply $\frac{R}{T}$, where R is the length of the content and T is the stream interval. Thus, small start-up latency may incur a much higher transmission bandwidth and cost. The DINA system may also face such a difficulty.

For the second difficulty, the users viewing a multicast stream cannot freely interrupt the stream because there are other viewers. Therefore, NVOD and QVOD systems cannot allow VCR-like interactivity such as pause, resume, rewind, slow motion, fast forward, and so on. These systems also hinder the introduction of new forms of interactive media to be deployed. In recent years, one popular approach to offer some form of VCR-like interactivity over NVOD and QVOD systems is to add a storage unit to the set top box (STB) so as to cache all the available content being broadcast. Such systems suffer from a higher system cost and operational problems like storage unit failure and management.

It can be realised that the prior art may fail to provide a solution to the existing problems in VOD systems. Specifically, current VOD systems may not be able to provide the clients/users with desired interactive functions with a short start-up time, while at the same time minimising the network load. Therefore, it is an object of this invention to resolve at least some of the problems at set forth in the prior art. As a minimum, it is an object of this invention to provide the public with a useful choice.

Summary of the invention

Accordingly, this invention provides, in the broad sense, a method and the corresponding system for transmitting data over a network to at least one client having a latency time to initiate transmission of said data to the client. The method of this invention includes the steps of:

- generating at least one of anti-latency data stream containing at least a leading portion of data for receipt by a client; and
- generating at least one interactive data stream containing at least a remaining portion of said data for the client to merge into after receiving at least a portion of an anti-latency data stream.

The anti-latency data streams and the interactive data streams may be generated by at least one anti-latency signal generator and at least one interactive signal generator, respectively.

It is another aspect of this invention to provide a method and the corresponding system for transmitting data over a network to at least one client including the step of fragmenting said data into K data segments each requiring a time T to transmit over the network, wherein each of the K data segments contains a head portion and a tail portion, and the head portion contain a portion of data of the tail portion of the immediate preceding segment to facilitate merging of the K data segments when received by the client.

The K data segments may be generated by a signal generator.

It is yet another aspect of this invention to provide a method and the corresponding system for transmitting data over a network to at least one client having a latency time to initiate transmission of said data to the client, including the steps of:

- generating at least one anti-latency data stream containing at least a leading portion of data for receipt by the client;
- pre-fetching the leading portion in the client as pre-fetched data; and

- generating at least one interactive data stream containing at least a remaining portion of said data for the client to merge into the leading portion.

5 This invention also provides a method and the corresponding system for transmitting data over a network to at least one client including the steps of generating a plurality of anti-latency data streams, in which the anti-latency data streams include:

- a leading data stream containing at least one leading segment of a leading portion of said data being repeated continuously within the leading data stream; and
- a plurality of finishing data streams, each of the finishing data streams:
 - containing at least the rest of the leading portion of said data; and
 - repeated continuously within said finishing data stream, and wherein each successive finishing data stream is staggered by an anti-latency time interval.

10 This invention further provides a method and the corresponding system for transmitting data over a network to at least one client. The method includes the steps of generating M anti-latency data streams from 1 to M , wherein an m^{th} anti-latency data stream has F_m segments, and F_m is an m^{th} Fibonacci number; and wherein said F_m segments are repeated continuously within the m^{th} anti-latency data stream.

15 It is yet another aspect of this invention to provide a method and the corresponding system for transmitting data over a network to at least one client, said data being fragmented into K segments each requiring a time T to transmit over the network. The method includes the steps of generating M anti-latency data streams containing 1 to K anti-latency data segments, wherein the anti-latency data segments are distributed in the M anti-latency data streams such that an k^{th} leading segment is repeated by an anti-latency time interval $\leq kT$ within the anti-latency data streams.

20 This invention further provides a method for receiving data being transmitted over a network to at least one client. The data to be transmitted is fragmented into K

segments each requiring a time T to transmit over the network. The data is divided into two batches of data streams, the anti-latency data streams include M anti-latency data streams, and the interactive data streams includes N interactive data streams. The method for receiving the data includes the steps of:

- 5 - raising a request for said data. The request may be raised by a processor of the client; and
- connecting the client to the M anti-latency data streams and receiving data in the M anti-latency data streams. The client or the receiver may connect to the anti-latency data streams by a connector.

10 This invention also provides a method and a corresponding system for receiving data being transmitted over a network to at least one client, wherein said data includes a leading portion and a remaining portion, and the remaining portion is transmitted by at least one interactive data stream including the steps of:

- 15 - pre-fetching the leading portion in the client as pre-fetched data, which is contained in the buffer of the client; and
- merging the pre-fetched data to the remaining portion by a processor.

20 Further embodiments and options of the above methods and systems will be described in the following sections, and may then be apparent to one skilled in the art after reading the description.

Brief description of the drawings

25 Preferred embodiments of the present invention will now be explained by way of example and with reference to the accompany drawings in which:

Figure 1 shows the data stream structure of a NVOD system.

Figure 2 shows the data stream structure of a QVOD system.

Figure 3 shows the overall system architecture of the data transmission system of this invention.

30 Figure 4 shows the data streams arrangement of Configuration 1 of the data transmission system of this invention.

Figure 5 shows the data streams arrangement of Configuration 2 of the data transmission system of this invention.

Figure 6 shows the data streams arrangement of Configuration 3 of the data transmission system of this invention. Note the difference in the arrangement of the Group II data streams comparing with Figures 4 & 5.

Figure 7 shows yet another Group I data streams arrangement of Configuration 3.

Figure 8 shows the data streams arrangement of Group I data streams of Configuration 4 of the data transmission system of this invention.

Figure 9 shows yet another arrangement of Group I data streams of Configuration 4 of the data transmission system of this invention.

Figure 10 shows one of the data streams arrangement of Configuration 5 of the data transmission system of this invention. The particular arrangement of Group I data streams shown in this figure combines Configurations 1 & 3.

Figure 11 shows the system configuration of a multicast data streams generator of the data transmission system of this invention.

Figure 12 shows the system configuration of receiver of the data transmission system of this invention.

Figure 13 shows the local storage versus transmission bandwidth trade-off relationship.

20 Detail Description of Preferred Embodiments

This invention is now described by ways of example with reference to the figures in the following sections. Even though some of them may be readily understandable to one skilled in the art, the following Table 1 shows the abbreviations or symbols used through the specification together with their meanings so that the
25 abbreviations or symbols may be easily referred to.

Abbreviation/Symbol	Meaning
VOD	Video-on-Demand
NVOD	Near Video-on-Demand
QVOD	Quasi Video-on-Demand
DINA	Distributed Interactive Network Architecture, as described in PCT applications nos. PCT/IB00/001857 & 1858
VCR	Video Cassette-Recorder
STB	Set-Top-Box
DDVR	Diskless Digital Video Recorder, the client of the system
IVOD	Instant Video-on-Demand, possible name of the system of this invention
<i>J</i>	no. of anti-latency data segments in an individual anti-latency data stream (in Configurations 1 to 3) or no. of data segments of the leading portion of the data to be transmitted (Configuration 4)
<i>K</i>	no. of data segments of the data to be transmitted
<i>M</i>	no. of anti-latency (Group I) data streams
<i>N</i>	no. of interactive (Group II) data streams
<i>Q</i>	amount of data to be transmitted
<i>R</i>	time required to transmit <i>Q</i> data over the network
<i>S</i>	amount of data in each data segment
<i>T</i>	time required to transmit each data segment over the network
<i>A</i>	no. of data streams in Group I(1) streams
<i>C</i>	no. of data segments in the data of Group I(1) streams
<i>B</i>	no. of data streams in Group I(2) streams
<i>D</i>	no. of data segments in the data of Group I(2) streams
<i>E</i>	no. of data segments in the coarse jump interval

Table 1. Abbreviations and Symbols Used

Although the following description refers to the data to be delivered as being video, it is expressly understood that data in other forms may also be delivered in the system of this invention, for example audio or software programs, or their combination. For instance, this invention may be used for deploying an operating system software to a large number of clients through a network upon request. Further, this invention may be utilised in data transmission systems handling a large amount of data with repetitive content, for instance in a video system bus of a computer handling many complicated but replicated 3D objects. Moreover, this invention may not be limited to the transmission of digital data only.

In this invention, a multi-stream multicasting technique is used to overcome the existing problems in VOD systems as described in the Background section. By

using this technique, the users are allowed VCR-like interactivity without the need to add a storage unit at the STB and caching all the content that may be viewed by the user on a daily basis.

5 Figure 3 shows the system configuration. The multicast streams are generated from a multicast server unit. The streams are multiplexed onto the physical media and distributed to the end users through a distribution network. At each user end, there is a set top box (STB), such as DDVR, that selects a multitude of streams for processing. By arranging the content to be carried on the streams in a desired manner (as shown
10 later in Figures 4 - 10), the start-up latency may be minimized while the users are provided with interactive functions. The DDVR should have sufficient bandwidth, buffer and processing capability to handle the multi-streams.

The data transmission system of this invention, which may be called an IVOD system, may look similar to the NVOD system. However, the IVOD and NVOD systems are differentiated by the following points:

1. how the content is put on the staggered streams,
2. how the staggered streams are generated,
3. how the DDVR selects and processes the multitude of staggered streams to restore the content.

The word "staggered" used above and throughout the specification in describing the data streams refers to the situation that each of the data streams begins transmission at different times. Therefore, two "frames" of two adjacent data streams,
25 in which the term "frame" represents the repeating unit of each data stream, are separated by a time interval.

In the broad sense, the data transmission method and system may be described as providing two groups of data streams Group I and II. Group I data streams, which
30 may be termed anti-latency data streams, may serve to reduce latency for starting-up the transmission of the required data. Group I data streams may be generated by at least one anti-latency signal generator. Group II data streams, which may be termed interactive data streams, may serve to provide the desired interactive functions to the users. Group II data streams may be generated by at least one interactive signal

generator. For the interactive functions provide by Group II data streams, this can be referred to the applicant's PCT applications nos. PCT/IB00/001857 & 1858, the contents of which are now incorporated as references therein. The operation of the interactive functions is not considered to be part of the invention in this application
5 and the details will not be further described here.

The operation of the IVOD system can best be illustrated by the following examples. Each of these examples is a valid IVOD system but they all differ in details with various tradeoffs. These examples only intend to show the working principles of
10 IVOD systems and are not meant to describe the only possible ways of IVOD operation.

In the following examples, the content to be transmitted having a total amount of data Q requires a total time R to be transmitted over the network. The content, for example, may be a movie. The Q data is broken up into K segments each having an amount of data S . Each data segment requires a time T to be transmitted over the network. Q and S may be in the unit of megabytes, while R and T are units of time. For the sake of convenience, the data segments of the Q data are labelled from 1 to K respectively. Therefore, $K = \frac{R}{T}$. The Q data may be divided into a leading portion
15 and a remaining portion. In most cases, the Group I anti-latency data streams may contain the leading portion only. The Group II interactive data streams may contain the remaining portion or the whole set of the Q data, and this may be a matter of design choice to be determined by the system manager.

It should be noted that the system may still work if the individual data segment contains different amounts of data than each other, provided that they all required a time T for transmission. This may be achievable by controlling the transmission rate of the individual data segment. However, individual data segments may be preferred to have same amount of data S for the sake of engineering convenience. On the other
20 hand, it may be relatively more difficult to implement the system for each of the data segments to have same amount of data S but with different transmission times.

Although the following description refers to the transmission of one set of data, for instance, a movie, it should be apparent to one skilled in the art that the method and system may also be adapted to transmit a certain number of sets of data depending on, for example, the bandwidth available.

5

A. Dual Streaming IVOD System (Configuration 1)

The simplest IVOD system is characterized by a dual-streaming operation. Dual streaming means that each user will tap into at most two of the multicast data streams at any time. Most of the time, the user may only be tapping into one data stream.

10

The segments are put onto the staggered streams as shown in Figure 4. There are two groups of staggered streams. For Group I anti-latency data streams, there are J segments on each frame. T is the anti-latency time interval and may also be the upper bound for the start-up latency of the IVOD system. Each anti-latency data stream is preferably staggered by the anti-latency time interval T , although the anti-latency time interval may be set at any desired value other than T .

15

In this particular example, J is equal to 16 and T is 30 seconds. So the frames in each of the Group I data streams repeat themselves after a time of JT being 8 minutes. There are a total of M streams in Group I.

20

For Group II interactive data streams, there are N interactive data streams, with each of them being staggered by an interactive time interval. Although the interactive time interval may again be set at any desired value, the interactive time interval is preferably to be JT (i.e. 8 minutes in this example) for the sake of engineering convenience. Assuming the length of the content is R (say R equals to 120 minutes), then there should be at least a total of $\frac{R}{JT} = 15$ streams in Group II. N may be larger than this value but this may create unnecessary network load.

25

30

When a user starts to view the content at time t_i , the DDVR at the user end will select one stream from Group I (Stream Ii) and one stream from Group II (Stream Iij) to tap into. Once the client connects to Streams Ii and/or Iij, the data streams are

processed by the DDVR, the client, and the segments are buffered according to the segment sequence number. The availability of the Group I staggered streams with stream interval T minimises the start-up latency to be equal to T .

Alternatively, the user or the client may tap into Stream I_i only and await all of the data in the leading portion to be received by the client before tapping into Stream I_j . After the DDVR has latched onto a Group I stream, the DDVR will immediately look for a suitable Group II stream for merging. In this particular case, each Group II data streams may preferably contain only the remaining portion of the Q data.

The method on merging of data streams can be found in the DINA technology. After merging, the Group I stream may no longer be needed and the DDVR may then rely solely on Stream I_j for subsequent viewing. This may be the optimised alternative only to minimise network load.

It should be noted that once the system has started, the user could initiate the following interactive requests, including pause and resume, rewind, and slow motion playback. However, forward and backward jumps may be restricted to jump to any one of the Group I or Group II streams (at any particular time). This problem may be resolved by fine-tuning the parameters of the system. For instance, Group I data streams may be designed to contain content that relatively few people wish to look at, like copyright notices.

The total number of streams in this type of IVOD system is $M + \frac{R}{JT}$. The

optimal system configuration is calculated to be $M = N = J = \sqrt{\frac{R}{T}}$, and the optimal total number of streams is given by $2\sqrt{\frac{R}{T}}$.

B. Dual Streaming IVOD System (Configuration 2)

The second example of IVOD system is also characterised by a dual-streaming operation. Again, the content is broken up into K segments of regular length T , and

the segments are labelled from 1 to K respectively. The segments are put onto the staggered streams in a pattern as shown in Figure 5.

In this configuration, there are also two groups of staggered streams. For Group I anti-latency data streams, there are J segments on each frame and the frames are repeated on each stream. In this example, J is again chosen to equal to 16 and T is 30 seconds. This configuration characterises in that one of the Group I data streams, Stream I1, contains only Segment 1 repeated in all time slots. Streams I2 to I9 contain Segment 2 to 17. In another words, Segment 1 may be viewed as a leading data stream containing the leading segment of the leading portion. Segments 2 to 9 may be considered as a plurality of finishing data streams containing the rest of the leading portion in the number of J segments. The Group I stream interval may be chosen to be any desired value, but is again preferably set to be T due to same reason as in Configuration 1. Streams I2 to I9 repeat themselves after JT (i.e. 8 minutes in this example).

In this particular example, there should be at least a total of $M = \frac{J}{2} + 1$ streams in Group I for the smooth merging of the leading data stream and the finishing data stream. M may be less than this value but then the user may suffer from the phenomenon of "dropping frames". M may be larger than this value but this may create unnecessary network load. This may be a matter of design choice that should be left to be determined by the system administrator.

Although the leading segment shown in Figure 5 contains only one leading segment, it should be understood that the leading data stream may contain more than one leading segment, for example, segments 1-4. The above conditions of the Group I anti-latency data streams of this Configuration 2 may then be viewed as T being four times as long, while this change may not affect the Group II interactive data streams. In such cases, the user may suffer from a larger start-up latency. On the other hand, M may be substantially reduced and could be $M = \frac{J}{8} + 1$ for the smooth merging of the leading data stream and the finishing data stream. Although this may be less

desirable, this may be again a matter of design choice that should be determined by the system administrator.

For Group II streams, the arrangement and the set up of the streams may be the same as in the previous example, and the same setting and variations is also applicable to this application.

When a user starts to view the content at time t_i , the DDVR at the user end will immediately tap onto Stream II. The start-up latency should be bounded to T as the leading segment is repeated every time period T . After all data in the leading segment is received, the DDVR will also tap onto one of the Group I finishing data streams, I2 to I9 in this case. For the ease of illustration, Stream Ii is chosen. As an alternative, the DDVR may tap onto the leading data stream and one of the finishing data streams simultaneously if the DDVR is capable of doing so. In the latter case, both streams are processed by the DDVR and the segments are buffered according to the segment sequence number.

The DDVR will also tap onto one of the Group II streams (in this case Stream II2). The time at which the DDVR taps onto the Group II streams is a matter of choice - it may do so:

1. immediately after tapping onto the leading data stream Stream II
2. immediately after tapping onto one of the finishing data streams
3. after all data in the leading portion contained in Group I data streams is received by the DDVR

Generally, the DDVR should tap onto one of the Group II streams at least right before all data in Group I streams is received or played by the client.

After all data in the Group I data streams has been buffered and received, the DDVR then merge onto one of the Group II streams. The merging technique is described in the DINA technology. After merging, the Group I stream (i.e. Stream Ii) may no longer be needed and the DDVR may rely only on the Group II stream for subsequent viewing to save bandwidth. Any allowable interactive request received at any time can be entertained as previously shown in the DINA technology.

The total number of streams in this IVOD system is $(\frac{J}{2}+1)+N$. As N preferably equals to $\frac{R}{JT}$, the optimal configuration is given by $J = \sqrt{2K} = \sqrt{\frac{2R}{T}}$ and the optimal total number of data streams of the system is equal to

5 $\sqrt{2K} + 1 = \sqrt{\frac{2R}{T}} + 1.$

C. Dual Streaming IVOD System (Configuration 3)

The third example of IVOD system is also characterised by a dual-streaming operation with the segments arranged in a hierarchical periodic frame structure with a size based on the Fibonacci numbers. Again, the content is broken up into K segments of regular length T , and the segments are labelled from 1 to N respectively. The segments are put onto the staggered streams in a pattern as shown in Figure 6. There are also two groups of staggered streams.

15 In this configuration, Group I data streams contains the data in the leading portion having J segments. Note that this J is slightly different from those used in Configurations 1 and 2. There are M Group I data streams labelled from 1 to M . For each of the Group I stream I_m , where m is an integer representing the stream number, the frame period is given by F_m where F_m is the m -th Fibonacci number. The first few

20 Fibonacci numbers are shown in Table 2. The Fibonacci numbers have the property that $F_y = F_{y-1} + F_{y-2}$, where y is an integer starting from 3. The Group I stream interval is again preferably set to be T as in Configurations 1 and 2. There are 12 Group I streams in this example. For Group II streams, the arrangement and the set up of the streams are similar to the previous examples, but for the sake of illustration, the

25 Group II streams starting at Segment 81.

j	1	2	3	4	5	6	7	8	9	10	11	12
Fj	1	2	3	5	8	13	21	34	55	89	144	233

Table 2. Fibonacci numbers.

5 The principle of operation can best be explained by the following even though many different variations are possible. When a user starts to view the content at time t , the DDVR at the user end will immediately tap onto two Group 1 data Streams I1 and I2. Both Segment 1 from Stream I1 and Segment 2 or 3 from Stream I2 will be buffered. Now there are two segments in the buffer, and Stream I2 has a frame size of 2, Stream I2 can be smoothly merged into using the methodology as described in the DINA technology. Thus, the startup latency should be bounded to T . After Segment 1 has been received, DDVR will tap onto Streams I2 and I3. Since there are only two segments in Stream I2, Segment 3 will either be buffered during the time when Segment 2 is being received, or Segment 3 will be available on Stream I2 immediately following Segment 2's completion. After both Segments 2 and 3 have been received out, the DDVR will tap onto Streams 3 and 4, and the process continues as before. Both streams are processed by the DDVR and the extra segments are buffered according to the segment sequence number.

20 In the above discussion, the DDVR is presumed to connect to the 1st and 2nd data streams for starting-up the movie such that the latency is bounded to be T . However, if the user wishes, he may choose to first tap onto the m^{th} and $(m+1)^{\text{th}}$ data streams, wherein m is any number larger than 1. The user can still view the content but may be suffering from larger latency. This may be preferred by some users who wish to skip the first few minutes of a movie, for example.

30 By constructing the frame period of the streams according to the Fibonacci number F_m , after Stream I_{m-1} has been received, the DDVR would have buffered at least $F_m = F_{m-1} + F_{m-2}$ time slots. Using the merging methodology as described in the DINA technology, Stream I_{m-1} can be smoothly merged into Stream I_m as the frame size of Stream I_m is exactly F_m .

It is noted that after m segments are received, exactly m more segments would have been buffered because of the dual streaming arrangement. The DDVR preferably begin to merge onto one of the Group II streams, at the very least to save bandwidth, once the number of segments buffered has exceeded the size of the Group II stream interval (in this case 80 segments are needed for an 8-minute Group II stream interval). After merging, the Group I stream (i.e. Stream Ii) may no longer be needed and the DDVR may rely only on the Group II stream for subsequent viewing. Any allowable interactive request received at any time can be entertained as described in the DINA technology.

There is no optimal parameter for this Configuration. To save bandwidth, there should be no Group II data stream. However, users may only be able to enjoy limited interactivity depending on how much of the data is received and buffered in the DDVR. Specifically, the user may perform pause, resume, rewind, slow motion, and backward jump, but the user may not be able to perform fast forward and forward jump functions.

The number of Group I data stream required, M , is determined by the number of Group II data streams, which is in turn to be determined manually according to various system factors. With a given start-up latency T , the total number of streams required in this IVOD system can be found by looking up the necessary frame size from a table containing the relevant Fibonacci numbers. The minimal number of data streams should be M such that $F_M \geq \frac{2K}{N}$ for the smooth merging between the individual Group I data streams. M may be less than this value but then the user may suffer from the phenomenon of "dropping frames". M may be larger than this value but this may create unnecessary network load. This may be a matter of design choice that should be left to be determined by the system administrator..

Using this technique, the start-up latency T can be as low as 6 seconds (with an average of 3 sec), with a Group II stream interval of 8 minutes. The total number of streams required for a 2-hour content can be as low as only 26.

An alternative arrangement for the Group I streams is shown in Figure 7. Note that the frame structure of the streams only follows the Fibonacci sequence after Stream 4.

5 D. Multi-Streaming IVOD System (Configuration 4)

The previous three examples show several possible implementations of the IVOD systems with dual-streaming. In fact, there are many more possible implementations of the IVOD system, each depending on a different arrangement of the segments in different streams, and on the maximum number of streams that the end user DDVR must simultaneously tap into and process. The above three examples are relatively simple to understand and implement, but the number of streams used are not optimal because of the restriction that only two maximum streams are tapped into and processed at any given time. In the current configuration, a multi-streaming IVOD system with the optimal number of streams is demonstrated.

15 This configuration is realized with the assumption that all the streams that carried the content are all tapped into and processed by the end user DDVR. Figure 8 shows a possible optimal arrangement of the initial thirty segments or so in various streams based on the harmonic series approach. The segments are labelled 1, 2, 3, ... etc... The necessary and sufficient condition for guaranteeing the start up latency to be bounded within one slot interval using only an optimal number of streams is that the placement of the segments should be done in such a way that Segment j (i.e. the j -th segment from the beginning of the leading portion) should be repeated in every j time slots or less, for all j from 1 to J . For example, Segment 1 should be repeated in every time slot in order that the start-up latency is bounded within one anti-latency interval T . Therefore, there may be a whole stream taken up by Segment 1 alone. Segment 2 should be repeated in every other time slot in order that the second segment is available immediately after the first segment has been received. Similarly Segment 3 should be repeated in every three time slots and Segment j should be repeated in every j time slots. For $j > 1$, the segment j may be repeated more frequently than required. That is, the j^{th} segment is repeated by an anti-latency time interval $\leq jT$. Note that the definition of the term "anti-latency time interval" in this Configuration 4 is different from that in Configurations 1 to 3.

The exact stream where the segments are placed does not matter as we are assuming that all streams are being received and processed by the DDVR. The segments are buffered by the DDVR and rearranged into a suitable order. The unfilled slots in Figure 9 can contain any data or even be left unfilled.

As in Configuration 3, there is no optimal parameter for this Configuration. To save bandwidth, there should be no Group II data stream, in which users may only be able to enjoy limited interactivity depending on how much of the data is received and buffered in the DDVR. This may not be desirable. The number of Group I data stream required, M , is determined by the number of Group II data streams, which is in turn to be determined manually according to various system factors. The total number M of streams required for carrying the J time slots can be

found by summing the harmonic series from 1 to J , such that $M \geq \sum_{j=1}^{j=J} \left(\frac{1}{j}\right)$. This

is approximately equal to $\gamma + \ln(J)$, where γ is the Euler's constant ($\sim 0.5772...$)

when J is large. Even though J can be set to any desired number larger than $\frac{K}{N}$, for

the sake of engineering convenience, it is preferred to have $J = \frac{K}{N}$, which equals to

the number of data segments in the interactive time interval. This is the optimal number of streams required to bind the start-up latency to within one slot interval.

To create an IVOD system based on this optimal multi-streaming condition, the streams are again divided into two groups, Groups I and II. The segment arrangements of the Group I streams has been shown in Figure 8. The segment arrangements of the Group II streams are same as those shown in any one of Figures 4 to 6. When a user initiates a viewing request, all of the Group I streams should be received and processed by the DDVR. In addition, a suitable Group II stream will also be tapped into and processed. This allows a smooth merging of the Group I streams (where the initial m segments are placed) into a single Group II stream. As an alternative, the tapping onto the Group II stream may await until all data in the leading portion contained in Group I streams is received by the client DDVR.

After one Group II stream interval (which is again set to be JT intentionally in this case), all the Group I streams may no longer be needed and only a single Group II stream is needed for the continuous viewing by the user. Like before, through the use of a plurality of Group II streams, once the system has started, the user could initiate any of the allowable interactive requests, including pause and resume, rewind, and slow motion playback.

As in configuration 3, it is possible to create an IVOD system entirely based on the group I streams as illustrated previously. By doing that, the number of streams can be reduced with minimised start-up latency. However, users of such systems may be restricted to limited interactivity, as discussed in Configuration 3. Furthermore, the buffer size at the DDVR must be as large as the entire content, and the processing capability of the DDVR is more demanding for the current configuration. The decision regarding which system to deploy should be left as an option to the service provider.

It should further be noted that this multi-streaming arrangement may be used to replace the Fibonacci stream sequences (Group I streams) in Configuration 4 to further reduce the number of streams required. The condition is that the DDVR should have enough buffer and processing power to buffer and process the received data. Table 3 in the up-coming section lists some results in all various configurations.

A non-optimal multi-streaming arrangement known as the logarithmic streaming is shown in Figure 9.

example). Furthermore, users can coarse jump at any time within the start-up period, the time when the DDVR connects to the Group I streams. The start-up period is preferably defined to be the time within the first Group II stream interval (that is, from the 0-minute point to the 9-minute point) as in previous configurations. Each coarse
 5 jump is 1 minute apart from each other, which is determined by the coarse-jump frame period. Thus, the users can skip the headers using this arrangement. The total number of streams needed for holding a two-hour content in the particular example shown in Figure 10 is 30.

10 Although Figure 10 only shows the combination of Configurations 3 and 1 in Group I data streams, it should be obvious to those skilled in the art that the following combinations are also possible:

- a. Configurations 4 and 1
- b. Configurations 3 and 2
- 15 c. Configurations 4 and 2

The number of Group I(1) data streams required, i.e. A , may be determined by taking E as $\frac{K}{N}$ in configurations 3 and 4. That is, if Configuration 3 is used in Group
 I(1), there should be A data streams in Group I(1) such that $F_A \geq 2E$. If

20 Configuration 4 is used, then $A \geq \sum_{c=1}^{c=C} \frac{1}{C}$. As in Configuration 4, C , the total number of data segments to be transmitted in Group I(1), preferably equals to E . The same considerations on the number of data streams required as in Configurations 3 and 4 may also be applicable to Group I(1).

25 The decision regarding which combination to deploy should again be left as an option to the service provider.

Additional Features of Individual Data Segments

30 To facilitate the change over of the streams without incurring substantial loss of data during the transition, the beginning of each data segment, which can be termed the head portion, may contain duplicated data appearing in the tail portion of the

immediate preceding segment. The amount of data to be carried in the duplicated portion may be T' (normalized with respect to the data rate of the stream), where T' is the delay that may incur during the change over of the streams. Typically, T' may be in the order of 10 - 20 milliseconds.

5

IVOD System Requirements

There are several system requirements:

- a. The server needs to generate the appropriate multi-streams in patterns that have been illustrated in any one of Configurations 1 to 5 or such patterns as may be designed.
- b. The distribution network should have sufficient capacity to carry all the required streams to the end user DDVR.
- c. The end user DDVR should have sufficient bandwidth, buffer and processing capability to handle the multi-streams. The DDVR should also have sufficient storage to buffer at least one Group II stream interval of data from the multi-streams.

10

15

These factors may affect the service provide in choosing which configuration to deploy.

20

Concept of Diskless DVR

Generally, the receiver DDVR may have a processor for raising request for the content, and a connector for connecting the Group I and II data streams.

25

For Configurations 1 and 2, it may be necessary for the DDVR to include a buffer for buffering the received Group I data streams. For Configurations 3 and 4, the DDVR should include a buffer for buffering the data received from Group I data streams. The processor will then also be responsible for processing the data to put the data in a proper order.

30

With the multi-streaming concept, the receiving device, the receiver, at the user end may not need to have any hard disk storage. The only memory or buffer needed at the STB, the client/receiver, may be the RAM (random-access memory) to buffer one stream interval equivalent of data. Assuming a stream interval of 8

minutes, this requires roughly 60 MB of RAM for a 1 Mb/s MPEG-4 stream. This technique can be contrasted with many VOD techniques that require a large hard disk storage (sometimes as large as 60 GB) at the STB. Therefore, this IVOD system also appears to the users like a diskless DVR. However, the system provider may choose to provide addition storage to the users in the form of hard disk or other non-volatile medium or use such other equipment as may be necessary to buffer and receive the data.

It should be further noted that there might be several options for the DDVR.

First, the DDVR may be configured such that it plays the received data at a slower rate than the transmission rate of the data. The transmission rate may be expressed in $\frac{S}{T}$ under the condition that each data segment contains same amount of data. In such cases, the DDVR may be required to have a larger buffer size to accommodate the un-received data.

Secondly, the DDVR may be configured to contain or pre-fetch at least a portion of the data in the Group I data streams, i.e. the leading portion of the data to be transmitted, for a certain period of time in its local buffer. Such data may be termed "pre-fetched data". If desired, the pre-fetched data may contain all of the data contained in the Group I data streams provided that the DDVR has adequate buffer size. In one extreme, the content of the data to be transmitted may be refreshed every day for video data, or more than once per day. In this particular example, it may be necessary for the pre-fetched data to be refreshed every day. The refresh time may be set at any desired value that may range from one day to even one year. It may be preferable to refresh the pre-fetched data during an off-peak period, like after midnight (for instance, from 01:00-06:00), or between 10:00 to 15:00, wherein the network activities resulting from clients' requests may be at a minimum. This process may be initiated by the anti-latency signal generator, the interactive signal generator, or by the client itself by a routine call procedure. In doing so, the latency time and the total number of data streams required in the network may be further reduced. This may be particularly important for VOD systems transmitting a large number of sets of data.

Trade-off of Space-Time-Bandwidth

There is a trade-off relationship for different configurations of the IVOD systems of this invention among buffer storage at DDVR (space), start-up latency
5 (time) and streams (transmission bandwidth) required. This is shown in Table 3 and further illustrated in Figure 13.

In Figure 13, the Vertex 1 may be realised as current VOD systems with all the data being sent and then stored in the STB, whether the client raises a request for the
10 data or not. In such a case, the STB should have a relatively large buffer size. This may increase the manufacturing costs of the STB.

Vertex 2 may represent the systems as described in Configurations 1-5. Under such a configuration, the requirement on the STB may be minimal while the system
15 may be more demanding on the bandwidth.

Vertex 3 may represent a hybrid system of Vertexes 1 and 2.

The decision on which "Vertex" to choose may be a matter of design choice
20 depending on various factors including the bandwidth available, the specification of the STB, local requirements on latency and interactivity, and so on.

Content Size L = 1 hr						
		Number of Streams Required				
Staggered Interval		6 min	7 min	8 min	10 min	15 min
Dual-Streaming	Configuration (1) T = 30 sec (coarse jump = 1 minute)	22	23	24	26	34
	Configuration (2) T = 30 sec (coarse jump = 2 minutes)	17	17	17	17	20
	Configuration (3) T = 6 sec (no coarse jump allowed)	20	19	18	17	16
	Configuration (5) T = 6 sec (coarse jump = 1 minute)	23	23	23	23	26
	Configuration (5) T = 6 sec (coarse jump = 2 minute)	22	22	21	20	21
Multi-Streaming Configuration (4)	Optimal Configuration T = 6 sec (no coarse jump allowed)	15	14	13	12	10
	Optimal Configuration T = 6 sec (coarse jump = 1 minute)	20	20	20	20	23
	Optimal Configuration T = 6 sec (coarse jump = 2 minute)	18	18	17	16	17
Content Size L = 2 hr						
		Number of Streams Required				
Staggered Interval		6 min	7 min	8 min	10 min	15 min
Dual-Streaming	Configuration (1) T = 30 sec (coarse jump = 1 minute)	32	31	31	32	38
	Configuration (2) T = 30 sec (coarse jump = 2 minute)	27	25	24	23	24
	Configuration (3) T = 6 sec (no coarse jump allowed)	30	27	26	23	20
	Configuration (5) T = 6 sec (coarse jump = 1 minute)	33	31	30	29	32
	Configuration (5) T = 6 sec (coarse jump = 2 minute)	32	30	28	26	25
Multi-Streaming Configuration (4)	Optimal Configuration T = 6 sec (no coarse jump allowed)	25	22	20	18	14
	Optimal Configuration T = 6 sec (coarse jump = 1 minute)	31	29	27	27	28
	Optimal Configuration T = 6 sec (coarse jump = 2 minute)	28	26	24	22	21

Table 3. Tradeoff among Buffer Storage (Space), Startup Latency (Time) and
Streams (Transmission Bandwidth) Required

Application to cable, satellite and terrestrial broadcasting systems

The IVOD systems of this invention may find immediate applications in existing cable TV, terrestrial broadcasting, and satellite broadcasting systems. With very little modification on the existing infrastructure, the non-interactive broadcasting, or NVOD systems may be converted into an IVOD system. Both analogue and digital transmission systems can take advantage of the multi-streaming concept. However, the discussions below will only describe system configurations for digital transmission systems.

10 In these digital broadcasting systems, the RF transmission bands are usually divided into 6 MHz (NTSC) or 8 MHz (PAL) channels. There can be over a hundred channels in cable TV, terrestrial or satellite broadcasting system. Figure 11 shows a typical system configuration for this multi-streaming system. It is very similar to existing broadcasting system. Only the transmission unit at the head end, which may
15 be called an anti-latency device, and reception unit at the user end, the client/receiver, may need to be modified. At the head end, instead of sending analog signals in each channel, digital signals such as QAM are transmitted. Typically, one can put in 30 - 40 Mb/s into an RF channel. Assuming a 2-hour content, one can first use MPEG-4 or other compression algorithms to convert the analog signal into a digital stream with a
20 bit rate of roughly 1 Mb/s. Using the Fibonacci dual-streaming (Configuration 3) or the optimal harmonic multi-streaming IVOD concept (Configuration 4), one can place 30 to 40 streams of the IVOD streams into a single RF channel. The contents are put into different RF channels according to the PAL / NTSC / SECAM standard to maintain compatibility with the existing broadcasting system, and each RF channel
25 can contain a few hours of contents.

At the user end, the set top box should be RF-tuned to the particular RF channel of interest. Then the cable modem would filter out the 30 - 40 Mb/s digital streams and decode two streams at a time (for Fibonacci dual-streaming systems) or
30 decode all the harmonic multi-streams (for harmonic multi-streaming systems). Figure 12 shows the block diagram of the STB / cable modem. The STB / cable modem is similar to other STB / cable modems except for its processing unit which can process at least 2 multi-streams simultaneously rather than a single stream. The decoded streams would be buffered in the STB and the content would be reconstructed

according to the sequence number of the segments. With the hundreds of channels available in a typical broadcasting system, this translates to over 200 hours or more of fully interactive programs available to an infinite number of users.

- 5 While the preferred embodiment of the present invention has been described in detail by the examples, it is apparent that modifications and adaptations of the present invention will occur to those skilled in the art. It is to be expressly understood, however, that such modifications and adaptations are within the scope of the present invention, as set forth in the following claims. Furthermore, the
- 10 embodiments of the present invention shall not be interpreted to be restricted by the examples or figures only.

20010731 16:40